Phylogenetic Postprocessing

Nicholas D. Pattengale, Bernard M.E. Moret
Department of Computer Science
University of New Mexico
Albuquerque, NM 87131, USA
{nickp,moret}@cs.unm.edu

1. Introduction

Phylogenetic reconstruction techniques often produce multiple, competing evolutionary hypotheses. The umbrella term *phylogenetic postprocessing* encompasses methods that attempt to reconcile the ambiguity. Three classes of phylogenetic postprocessing results are presented. (1) A sublinear $(1+\epsilon)$ approximation algorithm is derived for computing the familiar Robinson-Foulds (RF) distance [4] between two trees. (2) Standard consensus methods are augmented to take edge weight into consideration. A new consensus method based on edge weights is introduced. (3) A generalized family of metrics on tree space is derived. The metrics can be equipped with sensitivity to edge weights. Two members of the family are the RF metric and the weighted RF metric.

The time complexity of the RF approximation algorithm is logarithmic in the number of trees and completely independent of the size of each tree (save a more expensive, one time, embedding step). This algorithm is easy to implement and should prove particularly useful in clustering-based phylogenetic postprocessing because tree size is more prohibitive than the number of competing trees in phylogenetic reconstructions of biological datasets (as opposed to simulation-generated datasets).

The remainder of this extended abstract focuses solely on the RF approximation algorithm. The reader is referred to [3] and the poster in this conference for detailed treatment of the other results.

2. Background

2.1 Trees as Bit Vectors

We utilize an ability to represent phylogenetic trees unambiguously as vectors. We now outline the representation.

Denote the set of all possible unrooted, leaf-labeled trees on n taxa as $\mathcal{T}_n$. Notice that removing an edge in a phylogenetic tree splits the set of taxa in two. An edge is uniquely identified by the split that it induces. There are

$$b = \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \binom{n}{i} \approx 2^{n-1}$$

ways to split a set of taxa in two. By assigning indices to splits it is possible to represent trees as vectors. Denote the set of splits induced by the edges in tree T as $\Sigma(T)$. Assign indices to splits by using a function

$$f : \bigcup_{T \in \mathcal{T}_n} \Sigma(T) \rightarrow \mathbb{N}$$

which assigns each split to an integer on the interval $[1, b]$.

Definition 2.1. *The bit-vector representation of a phylogenetic tree T is $v_T \in \mathbb{R}^b$ where each element of v_T is taken as*

$$v_T[i] = \begin{cases} 1 & \text{if } f^{-1}(i) \in T \\ 0 & \text{otherwise} \end{cases}$$

Basically, for any tree, we construct a b-dimensional bit-vector such that bits are set based upon which splits are present in the tree.

2.2 Robinson-Foulds Metric

The usual way of comparing two trees is to count the number of edges in which they differ. This calculation defines the Robinson-Foulds (RF) metric [4].

$$d_{RF}(T_A, T_B) = \frac{1}{2}\left(|\Sigma(T_A) - \Sigma(T_B)|\right) + \frac{1}{2}\left(|\Sigma(T_B) - \Sigma(T_A)|\right)$$

where $-$ is set difference, $|.|$ is cardinality, and $+$ is arithmetic.

The bit-vector representation is well suited for calculating the RF metric. By construction, the RF metric is simply the $\|.\|_1$-norm (i.e. $\|v\|_1 = \sum_{i=1}^{b} |v[i]|$) of the difference vector between two trees in bit-vector form.

3. Sublinear Robinson-Foulds

We borrow a technique from geometry to compact vector dimensionality while preserving the Euclidian vector norm of difference vectors. We show that the technique yields a very good approximation of Robinson-Foulds distance while providing an asymptotic speedup over the standard method for computing RF distance.

3.1 The Embedding

Consider the following *randomized embedding* of a set of bit-vectors $\mathcal{V} \in \mathbb{R}^b$, $|\mathcal{V}| = m$. The technique is due to Indyk and Motwani [2].

Construct a $b \times \frac{4\ln(m)}{\epsilon^2}$ matrix, f, where each element is a random number taken from the Gaussian distribution with mean 0 and variance 1. Multiply the bit-vector representation of each tree by f. Now the new input set consists of m vectors each in $\frac{4\ln(m)}{\epsilon^2}$ dimensions. First notice that the dimensionality of each tree in the new input set is independent of original tree size and solely dependent on m, the number of trees.

The Johnson-Lindenstrauss lemma [1] can be used to prove that the $\|.\|_2$ norm of each embedded point $vf, \forall v \in \mathcal{V}$, obeys

$$(1-\epsilon)\|v\|_2 \leq \|vf\|_2 \leq (1+\epsilon)\|v\|_2$$

with probability $1 - \frac{1}{n^2}$.

In other words, the $\|.\|_2$ (Euclidian) norm between embedded vectors is an arbitrarily good approximation of the $\|.\|_2$ norm between non-embedded vectors. See figure 1 for a graphical representation of the embedding step.

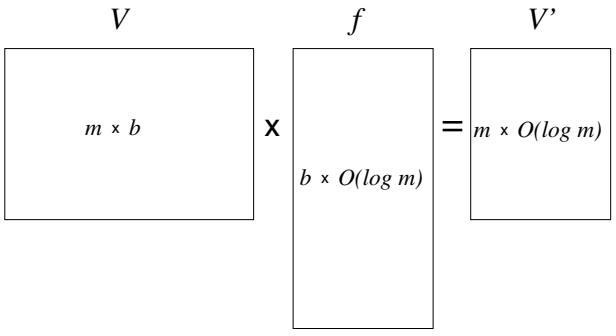

Figure 1. A sketch of randomized embedding. Each tree is a row in V. f is a random matrix with elements drawn from the Gaussian distribution. Each row of V' is the embedded representation of the corresponding row vector in V.

3.2 The Robinson-Foulds Connection

We take a set of m trees, convert them to bit-vector notation, and embed them into a space with $O(\log\,(m))$ dimensions as just outlined. As mentioned previously, the embedding preserves the $\|.\|_2$ norm between vectors. The RF distance between trees in bit-vector notation is the $\|.\|_1$ norm. However notice that by working with bit vectors (rather than vectors over a larger field) we have[3]

$$\|.\|_1 = (\|.\|_2)^2$$

.

Thus to calculate the $\|.\|_1$ norm of an arbitrary difference vector between two bit-vector trees it is sufficient to calculate the $\|.\|_2$ norm.

Accordingly, calculating an arbitrarily accurate approximation of RF distance amounts to vector difference followed by vector norm (inner product) on vectors in $O(\log(m))$ dimensions.

The prohibitive quantity in phylogenetics tends to be n, the number of leaves in a tree, rather than m, the number of plausible trees. Thus we anticipate that $O(\log\,(m)) << O(n)$ often, and that our technique will prove useful in practice.

4. Conclusion and Future Work

Since the embedding step itself has time complexity supra-logarithmically dependent upon m, our technique becomes useful when the number of pairwise RF calculations is large enough such that the complexity of the standard technique asymptotically exceeds the cost of embedding. This is typical in applications such as clustering. We plan to implement the technique and empirically assess its performance in such situations.

References

[1] P. Indyk. Algorithmic applications of low-distortion geometric embeddings. In *FOCS '01: Proceedings of the 42nd IEEE symposium on Foundations of Computer Science*, page 10. IEEE Computer Society, 2001.

[2] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings STOC 1998*, pages 604–613, 1998.

[3] N. D. Pattengale. Phylogenetic postprocessing. *Master's Thesis*, May 2005.

[4] D. Robinson and L. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.